CIVIC DIGITAL FELLOWSHIP

# Researching Alternative Machine Learning NAICS Classification Methods

**Abhinay Dommalapati**

Supervised by Justin Nguyen, Brian Dumbacher, and Daniel Whitehead

Economic Statistical Methods Division

# NAICS (North American Industry Classification System)

| 2015 Top NAICS Codes | |
|---|---|
| NAICS Code | NAICS Description |
| 336411 | AIRCRAFT MANUFACTURING |
| 541330 | ENGINEERING SERVICES |
| 541712 | RESEARCH AND DEVELOPMENT IN THE PHYSICAL, ENGINEERING, AND LIFE SCIENCES (EXCEPT BIOTECHNOLOGY) |
| 561210 | FACILITIES SUPPORT SERVICES |
| 236220 | COMMERCIAL AND INSTITUTIONAL BUILDING CONSTRUCTION |
| 541512 | COMPUTER SYSTEMS DESIGN SERVICES |
| 336611 | SHIP BUILDING AND REPAIRING |
| 541519 | OTHER COMPUTER RELATED SERVICES |
| 336414 | GUIDED MISSILE AND SPACE VEHICLE MANUFACTURING |
| 524114 | DIRECT HEALTH AND MEDICAL INSURANCE CARRIERS |
| 336413 | OTHER AIRCRAFT PARTS AND AUXILIARY EQUIPMENT MANUFACTURING |
| 541611 | ADMINISTRATIVE MANAGEMENT AND GENERAL MANAGEMENT CONSULTING SERVICES |
| 325412 | PHARMACEUTICAL PREPARATION MANUFACTURING |
| 334511 | SEARCH, DETECTION, NAVIGATION, GUIDANCE, AERONAUTICAL, AND NAUTICAL SYSTEM AND INSTRUMENT MANUFACTURING |
| 541710 | RESEARCH AND DEVELOPMENT IN THE PHYSICAL, ENGINEERING, AND LIFE SCIENCES |
| 541990 | ALL OTHER PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES |
| 324110 | PETROLEUM REFINERIES |
| 488190 | OTHER SUPPORT ACTIVITIES FOR AIR TRANSPORTATION |
| 541511 | CUSTOM COMPUTER PROGRAMMING SERVICES |
| 336412 | AIRCRAFT ENGINE AND ENGINE PARTS MANUFACTURING |
| 562910 | REMEDIATION SERVICES |
| 237990 | OTHER HEAVY AND CIVIL ENGINEERING CONSTRUCTION |
| 561612 | SECURITY GUARDS AND PATROL SERVICES |
| 336992 | MILITARY ARMORED VEHICLE, TANK, AND TANK COMPONENT MANUFACTURING |
| 517110 | WIRED TELECOMMUNICATIONS CARRIERS |

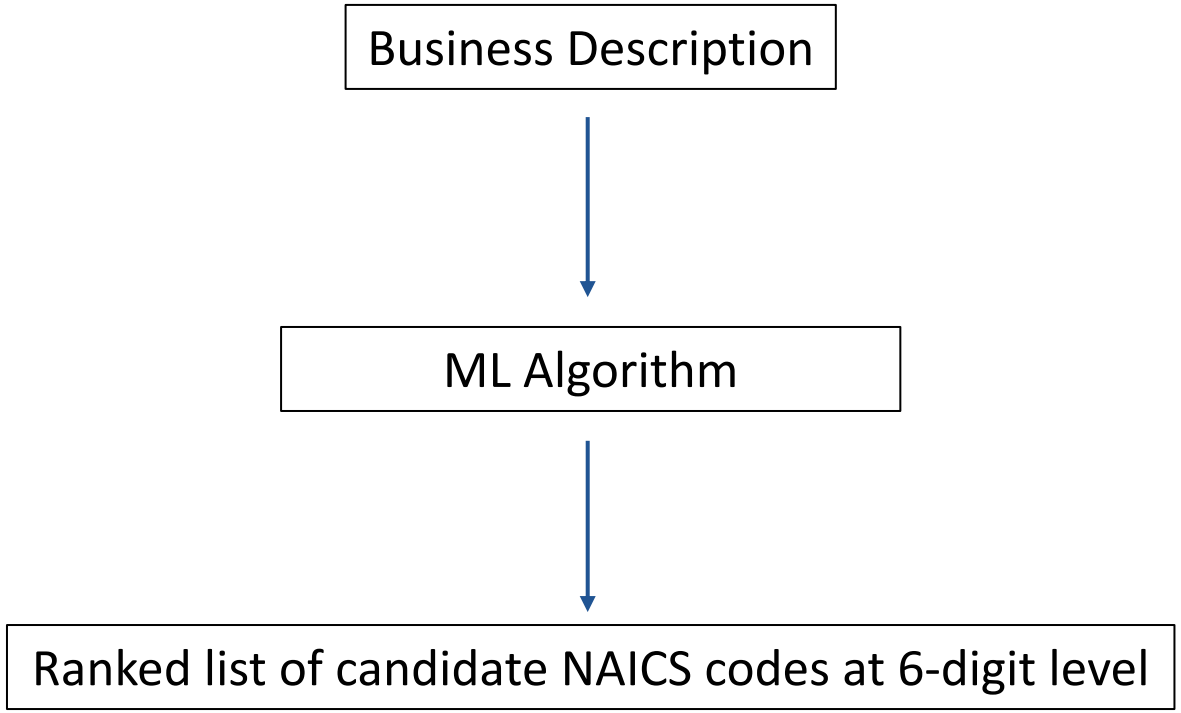
→ The standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy

# Research Problem and Goal

→ Some businesses are assigned the wrong NAICS code when respondents fill out an initial principal business activity (PBA) write-in

**Major Issues**

- **Time consuming**
- **Not accurate**
- **Difficult to use**



# Methodology

## Text Cleaning

- Convert to lowercase
- Manage numbers & punctuation
- Remove “stop” words
- Stem (ex. “manufacturing” → “manufactur”, “cars” → “car”)
- Consolidate (ex. “mfg” → “manufactur”, “automobile” → “car”)
- Create 0/1 variables that indicate the presence of words

## Current Methodology

- Hierarchal approach of first classifying the sector and then multiple models to drill down to 6-digit level
- Logistic Regression model that assigns probabilities to NAICS
- Information retrieval model that assigns relevance scores to NAICS
- ~76% accuracy

## Alternative Methodology

- Create numerical statistics intended to reflect how important a word is to a document in a collection or corpus (Term Frequency-Inverse Document Frequency)
- K-Nearest Neighbors model that stores all available classes and classifies new classes based on a similarity measure and assigns probabilities to NAICS

# Results from Alternative Methodology

- → Alternative methodology applied to NAICS 11 (Agriculture, Forestry, Fishing, and Hunting)
- → KNN model outputted ~90 - 98% accuracy from 4 TF-IDF measures
  - → Traditional TF-IDF
  - → Inter-class dispersion TF-IDF
  - → Modified IDF TF-IDF
  - → Class Frequency TF-IDF
- → Program ran in ~10-15 minutes

| | Precision (avg) | Recall (avg) | F1-Score (avg) |
|---|---|---|---|
| Traditional TF-IDF | 0.92 | 0.89 | 0.90 |
| W1 TF-IDF | 0.92 | 0.89 | 0.90 |
| W2 TF-IDF | 0.92 | 0.89 | 0.90 |
| W3 TF-IDF | 0.98 | 0.98 | 0.98 |

# Next Steps

- Apply alternative methodology to all other economic sectors and NAICS data as a whole
- Conduct model evaluation methods and compare with the Logistic Regression model/method
- Build out an interface for users (respondents, clerks, analysts, etc.)